ABSTRACT

Indexing documents used for searching documents or scientific publications. The performance of applications working with large databases depends on how quickly the database can process a request. The faster the database can process the indexed result become important. Along with the configuration that gave the most optimum result.

This research trying to connect two different infrastructures; Elasticsearch is an open source, distributed and nearly real-time search engine, based on Lucene. In the fields of cloud storage and cloud calculation, Elasticsearch is reliable, fast, and steady, and it can support the index of JSON data type by using HTTP; and The Hadoop Distributed File System (HDFS) is the primary storage system used by Hadoop applications. Methods used is; the data acquisition, data extraction using Apache Tika, setting the index structure, exporting the indexed file into csv, then processing the exported indexed file on the MapReduce Example programs that runs on the HDFS.

The results of the experiment show that, it is not possible to connect Elasticsearch directly with HDFS; compared to the non-indexed data file with the indexed data for the processes runs on the MapReduce Example program, it shown that the indexed one, have a significant process time than the non-indexed one; lastly, the greater number of mapper and reducer configuration does not guarantee the better performance of the process.

Keyword: indexing, Elasticsearch, extracting, Apache Tika, HDFS, Hadoop, Big data, mapper, reducer

ABSTRAK

Dokumen indexing banyak digunakan untuk proses pencarian dokumen ataupun publikasi sains. Indexing documents used for searching documents or scientific publications. Kinerja aplikasi yang bekerja dengan database besar bergantung pada seberapa cepat database dapat memproses permintaan. Semakin cepat database bisa mengolah hasil yang diindeks menjadi penting. Seiring dengan konfigurasi yang memberikan hasil paling optimal.

Penelitian ini mencoba menghubungkan dua infrastruktur yang berbeda; Elasticsearch adalah mesin pencari open source, terdistribusi dan hampir real-time, berdasarkan Lucene. Di bidang *cloud storage* dan *cloud calculation*, Elasticsearch dapat diandalkan, cepat, dan dapat mendukung indeks tipe data JSON dengan menggunakan HTTP; *Hadoop Distributed File System* (HDFS) adalah sistem penyimpanan utama yang digunakan oleh aplikasi Hadoop. Metode yang digunakan pada peneletian ini adalah; Akuisisi data, ekstraksi data menggunakan *Apache Tika*, pengaturan struktur indeks, mengekspor file yang diindeks ke dalam csv, kemudian memproses file yang telah diindeks dan kemudian diekspor ke *MapReduce example program* yang berjalan di HDFS.

Hasil percobaan menunjukkan bahwa, tidak mungkin menghubungkan *Elasticsearch* secara langsung dengan HDFS; Dibandingkan dengan file data yang tidak diindeks dengan data yang sudah diindeks untuk proses yang berjalan pada *MapReduce example program*, ditunjukkan bahwa yang diindeks, memiliki waktu proses yang signifikan daripada yang tidak diindeks; Terakhir, semakin banyak konfigurasi mapper dan reducer yang tidak menjamin kinerja proses yang lebih baik.

Keyword: indexing, Elasticsearch, extracting, *Apache Tika*, HDFS, Hadoop, Big data, mapper, reducer